From the Data to the Statistical Analysis of Football:
The Case of the Italian Serie A League

4

Alessio Drivet
Geogebra Institute of Turin, Italy

INTRODUCTION

This article deals with Data Analysis and Statistics applied to football. There are many publications dealing with this topic, but here the attempt is to explain how it can be attractive for secondary school students to learn notions of statistics from the analysis of data concerning the most popular of sports. In this case the analysis focuses on the last years of the Italian Serie A League.

BACKGROUND

The statistical analysis of Big Data is probably the most advanced frontier of sport and, more recently, of football. This is true even if we must remember that analyses on the dynamics of football have been present for many years (Grehaigne, 1997; Bouthier & David, 1997), and we cannot forget that this sport has found writers who, as fans and / or sportsmen, have revived, through their writings, the charm of this game (Hornby, 1992; Soriano, 1998).

We can basically identify two components that collect data:

1. sports clubs, through a network of observatories and the use of video and computer technologies;
2. specialized companies such as Opta, Prozone, StatDNA, Wyscout, etc., which, having created large databases, can provide services against payment.

These data serve as a basis for coaches to define modules and tactics and for technical managers to guide market choices.

As declared by Davide Nicola (Corriere della Sera of 1 April 2018, page 48), one of the Italian football coaches most involved in this approach: “Everything, or almost everything, is traceable to numbers, the analysis of big data allows you to discover links between the phenomena that happen during a match and therefore to predict the future ones”.

Finally, there is a further component made up of researchers studying the phenomenon and authors of books revealing the most relevant facts to the public. The fact that the soil is fertile is demonstrated by the large number of participants in events such as the Sports Analytics Conference in Boston or, as far as Italy is concerned, the Hackaton in Trento. The winning project of this Hackathon, organized by the FIGC (Italian Football Federation) on “match analysis” mixes “subjective” elements, for example the votes given by journalists, to numbers taken from statistical sources.

DOI: 10.4018/978-1-7998-3473-1.ch051

An approach based essentially on the analysis of objective data is, for example, the one known as the POGBA algorithm (Prediction of Goals by Assessing Phases) in which the available spatio-temporal data are used to evaluate the probability that a specific game situation will lead to an attempt of realization, and therefore to estimate the probability that the same attempt will lead to the expected goal (Decroos, Dzyuba, Van Haaren & Davis, 2017).

To conclude, we would like to mention an interesting point that is based on studies on artificial intelligence and neural networks; with this approach, we can segment game into sequences of situations that are discovered in an unsupervised way and we can learn *conceptors* (a mechanism of neurodynamical pattern learning and representation) that are useful for the prediction of the future of the match (Michael, Obst, Schmidsberger & Stolzenburg, 2017)

The problem is that all this is reserved for a group of specialists (Hendriks, 2016) while it would be interesting to be able to transfer part of these resources to an audience of students having two objectives:

1. explain how and where information can be found;
2. provide a concrete view of some statistical concepts and probabilities.

Since the aim is to deal with a level of information specific for the high school students, it will be used a type of statistics relatively simple: percentages, averages, probability distributions, regression and correlation.

For the first two concepts there is not much to say as they are part of a knowledge already present at previous levels of school.

As far as probability distributions are concerned, we shall limit ourselves to a simple reference.

The probability distribution is a model that associates a probability to each observable mode of a random variable.

We can distinguish two types of random variables, discrete or continuous: the probability distribution is discrete when the phenomenon is observable with an integer number of modes, and continuous when the random variable assumes a continuous set of values.

The text compares the discrete goal distribution with the Poisson distribution, used to calculate the probability P(x) of success of a repeated event n times in an experiment in a given unit of time. Indicating with $\lambda = np$ the average of the successes results:

$$P\left(x\right) = \lambda^{x} \frac{e^{-\lambda}}{x!}$$

The last two terms require, from an educational point of view, a clear distinction between function (already known) and regression. The first one indicates a precise relationship between the variables while the second one, used for economic, social, biological, studies etc., describes a phenomenon, except for a certain amount of variability that is not explained by the equation. This equation defines the curve that best approximates the data; it is called the regression curve of Y on X because the dependent variable is estimated by means of the independent one. Finally, it is possible to interpret the correlation from the point of view of regression using the coefficient of determination R^2. This coefficient represents the proportion of the variation explained by the independent variable. The square root of this value coincides with the linear correlation coefficient.